러프 집합에 기반한 불완전 정보의 정보 이론적 척도에 관한 연구

김국보[†] · 정구범[††] · 박경옥[†††]

요 약

러프집합에서는 식별불능관계와 근사공간 개념을 이용해서 불완전 정보로부터 최적화된 결정규칙을 유도하게 된다. 그러나, 처리하고자 하는 정보에 정량적이거나 중복 또는 누락된 데이터가 포함된 경우에는 오류가 발생될 수 있으므로, 이러한 데이터들을 제거하거나 최소화시키는 방법이 필요하다. 정보처리 분야에서 불확실성이나 정보의 양을 측정하는데 사용되고 있는 엔트로피는 러프 관계 데이터베이스의 불완전 정보를 제거하는데 사용되었다. 그러나, 정보시스템에 포함될 수 있는 불완전 정보를 모두 다루지는 못하였다.

본 논문에서는 정보시스템의 조건속성과 결정속성에 포함될 수 있는 불완전 정보를 제거하기 위한 정보 이론적 척도로서 러프집합을 이용한 객체관계 엔트로피와 속성관계 엔트로피를 제시한다.

The Study on Information-Theoretic Measures of Incomplete Information based on Rough Sets

Guk-Boh Kim[†], Gu-Beom Jeong[††] and Kyung-Ok Park[†††]

ABSTRACT

This paper comes to derive optimal decision rule from incomplete information using the concept of indiscernibility relation and approximation space in Rough set. As there may be some errors in case that processing information contains multiple or missing data, the method of removing or minimizing these data is required. Entropy which is used to measure uncertainty or quantity in information processing field is utilized to remove the incomplete information of rough relation database. But this paper does not always deal with the information system which may be contained incomplete information.

This paper is proposed object relation entropy and attribute relation entropy using Rough set as information theoretical measures in order to remove the incomplete information which may contain condition attribute and decision attribute of information system.

1. 서 론

Pawlak[5]에 의하여 제안된 러프집합(rough set) 이론은 정보시스템에서 상한 및 하한 근사(upper and lower approximation) 개념을 이용하여 불완전 정보를 취급하거나 불완전한 상태의 지식을 추론하기 위한 연역적 방법을 제공한다. 이러한 러프집합 이론은 불확실한 정보의 관리, 기계학습, 지식 발견, 부정확한 지식에 대한 표현 및 추론 등의 연구에 사용되고 있다[6,7].

정보시스템에서는 객체의 속성 값이 부정확하거나 누락된 경우 또는 복수 개의 값을 가지고 있는 등의 문제로 인하여 정보의 불완전성이 발생되므로, 일반적으로 데이터를 나타내는 [객체, 속성]이 정확하고 유일한 값을 갖는다고 가정함으로써 이러한 불완전성을 해결하였다[9]. 그러나, 실세계에서 [객체, 속성]은 항상 정확하고 유일한 값만을 가질 수가 없

[†] 종신회원, 대진대학교 컴퓨터공학과
[††] 정회원, 상주대학교 컴퓨터공학부
[†††] 삼성전자(주) 디지털프린팅(사) 선임연구원

으므로, 불완전성에 대한 문제 해결이 필요하다. 러프집합에서는 근사공간을 이용하여 데이터의 불완전성을 처리하였으나, 중복되거나 누락된 데이터의 처리등 모든 불완전 정보를 다루기에는 부족하다.

본 논문에서는 정보시스템에서 [객체, 속성] 값의 중복 및 누락으로 발생되는 불완전성을 해결하기 위한 정보 이론적 척도로서 러프 엔트로피(rough entropy)[1]의 확장된 개념인 객체 관계 엔트로피(object relation entropy)와 속성 관계 엔트로피(attribute relation entropy)를 제시한다.

2. 러프 집합과 러프 엔트로피

2.1 러프 집합

정보시스템 $S = \{U, A, V\}$라 하자. 객체들의 유한집합 $U = \{x_1, x_2, \cdots, x_n\}$, $U \neq 0$ 이며, A는 기본속성들의 유한집합이 된다. A에 있는 속성들은 조건속성 C와 결정속성 D로 분류되며, $A = C \cup D$, $V = \cup_{p \in A} V_P$, 이고, V_P는 기본속성 P의 영역이 된다.

속성들의 모든 부분집합 $P(P \subseteq A)$와 임의의 원소 $x_i, x_j \in U$ 라 하면, 식별 불가능 관계(indiscernibility relation)인 이진관계 $IND(P)$는 다음과 같이 정의된다.

$$IND(P) = \{(x_i, x_j) \in U \times U : \forall p \in P, \quad p(x_i) = p(x_j)\} \quad (1)$$

여기서, x_i, x_j 는 정보시스템 S 에서 속성 P의 집합에 의하여 식별 불가능하다고 말한다. 그리고 $P(x)$는 객체 x에 할당된 속성 P의 값으로서, $IND(P)$는 모든 $P \subseteq A$에 대하여 U에서 식별 불가능한 동치관계(equivalence relation)가 되며, 다음과 같은 관계가 성립된다.

$$IND(P) = \cap_{p \in P} IND(p) \quad (2)$$

정보시스템 $S = \{U, A, V\}$이고, $R \subseteq A$ 가 동치관계라면, 순서쌍 $AS = (U, R)$을 근사공간이라 한다. U의 원소 x_i에 대하여 $IND(P)$에서 x_i의 동치 클래스는 다음과 같다.

$$[x_i]_{IND(P)} = \cap_{R \in P}[X]_R \quad (3)$$

$X \subseteq U$ 라 하면, AS에서 X의 상한근사 R^*X와 하한근사 R_*X는 다음과 같다.

$$R^*X = \{ x_i \in U \mid [x]_R \cap X \neq 0 \} \quad (4)$$
$$R_*X = \{ x_i \in U \mid [x]_R \subseteq X \}$$

집합 $BN_R(X) = R^*X - R_*X$ 를 X의 R-경계(R-boundary)라고 한다.

집합 X에 대한 정확성 척도는 다음과 같다.

$$\alpha_R(X) = \frac{card\ R_*X}{card\ R^*X}, \quad X \neq 0 \quad (5)$$

여기서, $0 \leq \alpha_R(X) \leq 1$이 된다.

집합 X에 대한 지식의 불완전성의 정도를 나타내는 부정확성 척도는 다음과 같다.

$$\rho_R(X) = 1 - \alpha_R(X) \quad (6)$$

$\rho_R(X)$는 러프집합의 경계영역으로부터 발생되는 불완전성을 알기 위한 좋은 방법이지만, 부정확성 척도를 계산하기 위해서는 각 근사공간에 포함되어 있는 원소들에 대한 사전 지식이 요구된다. 또한, 근사공간에 포함되어 있는 원소들의 식별불능 관계에 대한 정보를 충분히 제공해 주지 못하는 문제점도 존재한다.

2.2 러프 엔트로피

통신이론을 위하여 사용된 샤논-엔트로피[8]는 다음과 같다.

$$H = -\sum_i p_i \log_2 p_i \quad (7)$$

엔트로피는 데이터베이스, 정보처리 및 의사결정 분야 등에서 불확실성이나 정보의 양을 측정하는 척도로 사용되고 있으며, 주로 퍼지집합에서 광범위하게 연구되고 있다. Beaubouef, Perty 및 Arora[1]는 러프 관계 데이터베이스에서 불확실성에 대한 정보-이론적 척도로 러프 엔트로피를 사용하였다. 러프 엔트로피는 전형적인 정보이론에서 엔트로피를 측정하기 위하여 사용된 샤논-엔트로피와 같은 방법으로 유도되었으며, 러프집합 X의 엔트로피 $Er(X)$는 다음과 같다

$$Er(X) = -(\rho_R(X))[\sum Q_i \log(P_i)] \quad (8)$$

여기서, $i = 1, \cdots, n$까지의 동치 클래스이다. $\rho_R(X)$는

러프집합 X의 부정확성 척도이며, $\sum Q_i \log(P_i)$는 러프집합 X의 전체 또는 부분에 속하는 각 동치 클래스에 대한 확률의 합이 된다. 만일 c_i가 동치클래스 i에 있는 원소들의 수이고 주어진 동치 클래스의 모든 멤버(member)가 같다면, 클래스 i에서 특정 속성 값이 존재할 확률은 $P_i = 1/c_i$이 된다. Q_i는 전체 집합에서 동치 클래스 i에 대한 확률로서 동치 클래스 i에 있는 원소들의 수를 모든 동치 클래스들의 전체 원소들의 수로 나눈 결과가 된다.

[예제] 러프 엔트로피의 예를 들기 위하여 러프집합 X의 하한 및 상한근사와 이들의 근사공간에서 분할된 동치관계를 다음과 같이 가정한다.

$R_* = \{1, 2, 3, 4\}$
$R^* = \{1, 2, 3, 4, 5, 6, 7\}$
$IND/E_1 = \{\{1, 2, 3, 4\}, \{5, 6, 7\}\}$
$IND/E_2 = \{\{1, 2\}, \{3, 4\}, \{5, 6, 7\}\}$
$IND/E_3 = \{\{1\}, \{2\}, \{3\}, \{4\}, \{5, 6, 7\}\}$

- 러프집합 X의 부정확성 척도는 $\rho_R(X) = 1 - (4/7) = 3/7$로서, 동치관계 E_1, E_2, E_3에서도 동일한 값이 되므로 각 클래스의 부정확성에 대한 차이를 구별할 수 없다.
- 러프 엔트로피 $Er(X)$는 동치관계 E_1, E_2, E_3에 대하여 다음과 같이 계산된다.

$$Er(E_1) = -(3/7)[(4/7)\log(1/4) + (3/7)\log(1/3) = 0.235$$
$$Er(E_2) = -(3/7)[(2/7)\log(1/2) + (2/7)\log(1/2) + (3/7)\log(1/3)] = 0.021$$
$$Er(E_3) = -(3/7)[(1/7)\log(1) + (1/7)\log(1) + (1/7)\log(1) = 0.019$$

계산 결과 러프 엔트로피는 $Er(E_1) > Er(E_2) > Er(E_3)$가 된다. 따라서, $Er(E_3)$의 불확실성이 가장 작으며, 러프집합의 부정확성 척도와는 달리 동치관계 E_1, E_2, E_3로 분할된 클래스의 원소가 모두 같더라도 동치관계의 멤버에 따라 엔트로피가 달라진다는 것을 알 수 있다.

러프 엔트로피는 정보시스템에서 누락되거나 부정확한 속성 값의 불확실성을 측정할 수 있기 때문에 정량화된 정보 내용의 불완전성을 제거하는데 매우 유용하지만, 속성 및 속성간의 식별불능 관계만을 고려하기 때문에 객체에 대한 불확실성의 척도로서는 미흡한 점이 있다.

3. 객체 관계 엔트로피와 속성 관계 엔트로피

3.1 불완전 정보시스템

정보시스템 $IS = \{U, A\}$에서 속성들의 집합 $A = C \cup \{d\}$라 하면, $d \notin C$를 결정속성이라 하고, $a \in C$에 대하여 $a : U \rightarrow V_a$를 만족하는 C를 조건속성이라 한다. V_a는 속성 값의 집합이 된다. 객체 $(x, y) \in U^2$, $\forall a, d \in A$일 때 다음 조건에 해당되면 불완전 정보시스템이 된다.

$$(a(x) = a(y)) \wedge (d(x) \neq d(y)), \text{ 또는} \quad (9)$$
$$(a(x) = \Lambda) \vee (a(y) = \Lambda) \vee (d(x) = \Lambda) \vee (d(y) = \Lambda)$$

여기서, Λ는 null을 나타낸다.

Beaubouef와 Kryszkiewicz[3]는 불완전 정보시스템의 범위를 불완전 조건속성인 $(a(x) = a(y) \vee a(x) = \Lambda \vee a(y) = \Lambda)$로 제한하였다. Beaubouef는 속성의 동치관계에 대한 불완전성만을 고려하였으며, 불완전 정보시스템의 원시 정보를 변경하지 않으면서 지식감축을 이용하여 불완전한 정보를 제거하는 방법을 사용하였다.

이 방법에서 $(a(x) = a(y) \wedge d(x) \neq d(y))$와 같이 객체의 조건속성 값은 모두 같으나 결정속성 값이 다른 경우와 $(d(x) = \Lambda \vee d(y) = \Lambda)$와 같이 결정속성 값을 알 수 없는 경우는 제외되었다. 그러나, 실제로는 조건속성에서 발생될 수 있는 불완전성은 결정속성에서도 동일하게 발생될 수 있으므로 결정속성의 불완전성에 대한 처리방법도 고려되어야 한다. 따라서, 이러한 문제의 해결 방안으로 Beaubouef의 러프 엔트로피를 기반으로 하는 객체 관계 엔트로피와 속성 관계 엔트로피를 제시한다.

3.2 객체 관계 엔트로피

러프집합 엔트로피 $Er(X)$를 보완한 객체관계 엔트로피는 null인 결정속성 값을 해당 객체의 전체 속성에 대한 엔트로피를 계산함으로서 결정속성의 값을 결정할 수 있으며, 조건속성 값은 동일하나 결정

속성 값이 다른 불일치 결정속성 값도 결정할 수 있다. 객체 관계 엔트로피는 다음과 같이 정의된다.

[정의 1] 러프집합의 객체 $x \in U$에 대한 객체 관계 엔트로피 $Eo(x_t)$는 다음과 같다.

$$Eo(x_t) = -\sum_j (\rho_j(R)) [P_i \log_2(P_i)] \quad (10)$$

여기서, $i = 1,\cdots,n;\ j = 1,\cdots,m;\ t = 1,\cdots,l$이다.

■ (x_t, d)의 결정속성 값 $a_d(x_t)$가 null이거나 또는 불일치 값(조건속성의 값이 동일하나 결정속성의 값이 다른 경우)일 때 대체되는 결정속성 값은 객체 관계 엔트로피의 정의에 따라 결정한다.

(1) 가능한 결정속성 값($a_d(x_1) \sim a_d(x_{l-t})$에서 null 및 동일한 속성 값 제외)을 차례로 $a_d(x_t)$에 적용하여 $Eo(x_t^1) \sim Eo(x_t^{l-t})$를 구한다. 여기서 $Eo(x_t^1) \sim Eo(x_t^{l-t})$는 null 값을 대체할 수 있는 결정속성 값인 객체 관계 엔트로피가 된다.

(2) $Eo(x) = \min\{Eo(x_t^1), Eo(x_t^2), \cdots, Eo(x_t^{l-t})\}$에 의하여 결정된 값으로 결정 속성 $a_d(x_t)$의 null 값 또는 불일치 값을 대체한다.

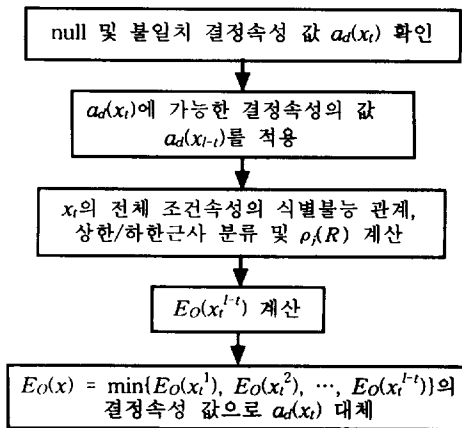



그림 1. 객체 관계 엔트로피의 수행 절차

3.3 속성 관계 엔트로피

속성 관계 엔트로피는 객체 관계 엔트로피 $Eo(x_t)$에 기반을 두고 있으며, 조건속성 값이 null일 경우에 해당 객체의 null 조건속성 값만 계산함으로서 효과적으로 null 값을 대체할 수 있게 한다.

[정의 2] 러프집합의 객체 $x \in U$에 대한 속성관계 엔트로피 $E_A(x_t)$는 다음과 같이 정의한다.

$$E_A(x_t) = (\rho_j(R)) [P_i \log_2(P_i)] \quad (11)$$

여기서, $i = 1,\cdots,n;\ j = 1,\cdots,m;\ t = 1,\cdots,l$이다.

■ (x_t, A_k)의 조건속성 값 $a_k(x_t)$가 null 값일 때 대체되는 조건속성 값은 속성 관계 엔트로피를 이용하여 다음과 같이 결정한다.

(1) 조건속성 A_k의 가능한 값($a_k(x_1) \sim a_k(x_{l-t})$에서 null 및 동일한 속성 값 제외)을 차례로 $a_k(x_t)$에 적용하여 $E_A(x_t^1) \sim E_A(x_t^{l-t})$를 구한다. 여기서 $E_A(x_t^1) \sim E_A(x_t^{l-t})$는 null 값을 대체할 수 있는 조건속성 값인 속성 관계 엔트로피가 된다.

(2) $E_A(x) = \min\{E_A(x_t^1), E_A(x_t^2), \cdots, E_A(x_t^{l-t})\}$에 의하여 결정된 값으로 결정 속성 $a_k(x_t)$의 null 값을 대체한다.

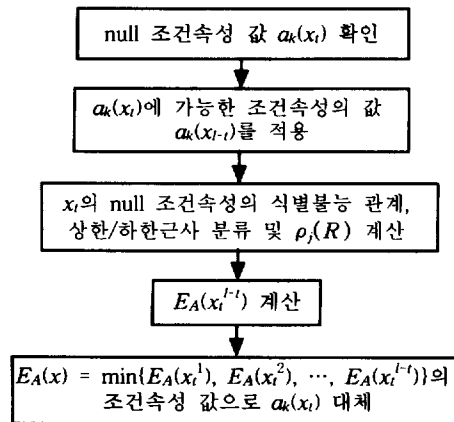



그림 2. 속성 관계 엔트로피의 처리 절차

3.4 러프 엔트로피 알고리즘

러프 엔트로피 알고리즘은 객체 관계 엔트로피와 속성 관계 엔트로피의 처리절차를 결합한 것으로, 정보시스템을 분석해서 조건속성과 결정속성의 불완전성에 따라 객체 관계 엔트로피와 속성 관계 엔트로피를 구한다. 즉, 조건속성의 null 속성 값에 대한 엔트로피를 계산하거나 또는 결정속성의 null 및 불일치 속성 값에 대한 엔트로피를 계산한다. 그리고, 엔트로피 계산 결과를 비교하여 해당 불완전 속성 값을 최적의 속성 값으로 대체시키는 속성 값을 결정하는 기능을 수행한다.

```
Algorithm Rough_Entropy
input   : Incomplete_Information
            - Null_Condition_Attribute_Value
            - Null/Discord_Decision_Attribute_Value
            - Multiple_Condition_Attribute_Value
output  : complete_Information
process :
  1. read Input_Information from database;
  2. [Find Incomplete_Information]
     while Incomplete_Information until not_exist

        if Condition_Attribute a_k(x_t) =
                         Null or Multiple_Value
              call step_3;
        else if  Decision_Attribute a_d(x_t) =
                            Null or Discord_Value
              call step_4;
     endwhile;
     go to step_5;

  3. [Attribute_Relation_Entropy_Module]
     if a_k(x_t) = Null
        for i = 1 to (l - t)
            a_k(x_t) =  a_k(x_i);
            compute ρ_j(R) using (Formula_6);
            compute E_A(x_t^i) using (Formula_11);
        endfor;
        E_A(x) = min{E_A(x_t^i)};
        a_k(x_t) = E_A(x);
     endif;
     if a_k(x_t) = Multiple_Value
        compute ρ_j(R) using (Formula_6);
        for i = 1 to l
            a_k(x_t) =  a_k(x_i);
            compute E_A(x_t^i) using (Formula_11);
        endfor;
        E_A(x) = min{E_A(x_t^i)};
        a_k(x_t) = E_A(x);
     endif;

  4. [Object_Relation_Entropy_Module]
     if a_d(x_t) =  Null or Discord_Value
        for i = 1 to (l - t)
           for j = 1 to m
              a_d(x_t) =  a_k(x_i)
              compute ρ_j(R) using (Formula_6);
              compute  E_O(x_t^i) using (Formula_10);
           endfor;
        endfor;
        E_O(x) = min{E_O(x_t^i)};
        a_d(x_t) = E_O(x);
     endif;

  5. write complete_information to database;
     endalgorithm;
```

4. 적용 사례

4.1 문제 설정

어떤 기업체의 신입사원 채용 여부를 판단하는 정보시스템을 표 1과 같이 구성하였다. 여기서 x_i는 신입사원 대상자이며, 조건 속성에 해당하는 평가 조건은 대학성적, 사회성 및 창의성의 3개 영역으로 구분하였다. 그리고, 결정 속성에 해당하는 판단결과는 평가 조건에 따라 결정되었다.

표 1. 신입사원 평가 테이블

구분	성적 (a)	사회성 (b)	창의성 (c)	결과 (d)
x_1	3.7	미흡	미흡	불합격
x_2	3.5	우수	보통	합격
x_3	3.0	보통	미흡	불합격
x_4	4.3	우수	우수	합격
x_5	3.6	우수	우수	합격
x_6	2.8	미흡	보통	불합격
x_7	4.0	우수	우수	합격
x_8	4.2	보통	보통	합격
x_9	3.4	미흡	미흡	불합격
x_{10}	3.3	보통	미흡	불합격

4.2 객체 관계 엔트로피의 적용

표 1의 신입사원 평가 테이블을 불완전 정보시스템으로 구성하기 위해서 표 2와 같이 테이블의 각 데이터를 코드화하고 불완전 정보를 포함시켰다. 표 2에서 조건속성은 {a, b, c} 항목이고 결정속성은 {d} 항목이며, 객체는 $\{x_1, x_2, \cdots, x_{10}\}$ 항목이 된다. 그리고, 결정속성 $\{x_{10}, d\}$의 속성 값에 Λ(null)이 포함되어 있는 불완전 정보시스템으로 가정한다면, $\{x_{10}, d\}$의 null 값을 대체할 수 있는 결정속성 값을 Rough_Entropy 알고리즘의 [Object_Relation_Entropy_Module]을 사용해서 구하게 된다.

결정속성 {d}에서 가능한 결정속성 값 $a_d(x_i)$의 유형은 {'1', '2'}이므로 $E_{O1}(x_{10})$과 $E_{O2}(x_{10})$의 엔트로피 $E_O(x_t^i)$를 계산한 다음, 그 결과를 비교하여 $\{d_{10}\}$의 값 $E_O(x)$을 결정한다. 단, 상한 및 하한근사를 구하는 것은 집합 X와 식별불능 관계 IND에 의하여 쉽게 산출할 수 있으므로 생략한다.

표 2. null 결정속성 값 포함

U	a	b	c	d
x_1	2	3	3	2
x_2	2	1	2	1
x_3	3	2	3	2
x_4	1	1	1	1
x_5	2	1	1	1
x_6	3	3	2	2
x_7	1	1	1	1
x_8	1	2	2	1
x_9	3	3	3	2
x_{10}	3	2	3	Λ

■ 성적(a)
1 → 4.0 이상
2 → 3.5~3.9
3 → 3.5 미만

■ 사회성, 창의성(b,c)
1 → 우수
2 → 보통
3 → 미흡

■ 결과(d)
1 → 합격
2 → 불합격

(1) $a_d(x_i)$ = '1'을 적용한 경우

$X_1 = \{x_2, x_4, x_5, x_7, x_8, x_{10}\}$

$IND/a = \{[x_4,x_7,x_8],[x_1,x_2,x_5],[x_3,x_6,x_9,x_{10}]\}$

$IND/b = \{[x_2,x_4,x_5,x_7],[x_3,x_8,x_{10}],[x_1,x_6,x_9]\}$

$IND/c = \{[x_4,x_5,x_7],[x_2,x_6,x_8],[x_1,x_3,x_9,x_{10}]\}$

$E_{O1}(x_{10}) = (1 - 3/10)[(4/10)\log_2(4/10)] + (1 - 4/10)[(3/10)\log_2(3/10)] + (1 - 3/10)[(4/10)\log_2(4/10)] = 1.053$

(2) $a_d(x_i)$ = '2'를 적용한 경우

$X_2 = \{x_1, x_3, x_6, x_9, x_{10}\}$

$IND/a = \{[x_4,x_7,x_8],[x_1,x_2,x_5],[x_3,x_6,x_9,x_{10}]\}$

$IND/b = \{[x_2,x_4,x_5,x_7],[x_3,x_8,x_{10}],[x_1,x_6,x_9]\}$

$IND/c = \{[x_4,x_5,x_7],[x_2,x_6,x_8],[x_1,x_3,x_9,x_{10}]\}$

$E_{O2}(x_{10}) = (1 - 4/7)[(4/10)\log_2(4/10)] + (1 - 3/6)[(3/10)\log_2(3/10)] + (1 - 4/7)[(4/10)\log_2(4/10)] = 0.858$

객체 관계 엔트로피의 계산 결과 $E_O(x) = \min\{E_{O1}(x_{10}), E_{O2}(x_{10})\}$이므로 $\{d_{10}\}$의 속성값 $a_d(x_{10})$은 $E_{O2}(x_{10})$에서 계산된 속성 값 '2'로 결정한다. $\{d_{10}\}$의 대체 속성 값의 정확성 여부는 표 1과 비교해 보면 알 수 있다. 즉, 두 테이블의 속성 값이 모두 '2'이므로, 객체 관계 엔트로피의 수식과 계산 결과는 정확하다고 볼 수 있다.

4.3 속성 관계 엔트로피의 적용

다음 결정 테이블은 표 2에서 사용된 테이블과 같으나, 조건속성 $\{x_9, b\}$의 속성 값에 Λ(null)이 포함되어 있다고 가정한다면, $\{x_9, b\}$의 가능한 조건속성 값을 Rough_Entropy 알고리즘의 [Attribute_Relation_Entropy_Module]을 사용해서 구하게 된다.

표 3. null 조건속성 값 포함

U	a	b	c	d
x_1	2	3	3	2
x_2	2	1	2	1
x_3	3	2	3	2
x_4	1	1	1	1
x_5	2	1	1	1
x_6	3	3	2	2
x_7	1	1	1	1
x_8	1	2	2	1
x_9	3	Λ	3	2
x_{10}	3	2	3	2

조건속성 {b}에서 가능한 조건속성 값 $a_k(x_i)$의 유형은 {'1', '2', '3'}이므로 $E_{A1}(x_9)$, $E_{A2}(x_9)$ 및 $E_{A3}(x_9)$의 엔트로피를 계산한 다음, 그 결과를 비교하여 $\{b_9\}$를 결정한다. 계산 과정은 4.2의 객체 관계 엔트로피 계산과 유사하므로 결과만 기술한다.

(1) $a_k(x_i)$ = '1'을 적용한 경우

$E_{A1}(x_9) = (1 - 2/10)[5/10)\log_2(5/10)] = 0.400$

(2) $a_k(x_i)$ = '2'를 적용한 경우

$E_{A2}(x_9) = (1 - 2/6)[(4/10)\log_2(4/10)] = 0.353$

(3) $a_k(x_i)$ = '3'을 적용한 경우

$E_{A3}(x_9) = (1 - 3/6)[(3/10)\log_2(3/10)] = 0.261$

속성 관계 엔트로피의 계산 결과 $E_A(x) = \min\{E_{A1}(x_9), E_{A2}(x_9), E_{A3}(x_9)\}$이므로, $\{b_9\}$의 속성 값 $a_k(x_9)$은 $E_{A3}(x_9)$에서 계산된 속성 값 '3'으로 결정한다. $\{b_9\}$의 대체 속성 값의 정확성 여부는 표 1과 비교

해 보면 알 수 있다. 즉, 두 테이블의 속성 값이 모두 '3'이므로, 속성 관계 엔트로피의 수식과 계산 결과는 정확하다고 볼 수 있다.

5. 결 론

러프 집합의 부정확성 척도가 불완전 정보시스템의 정확성을 향상시키지 못하는 반면, 러프 집합 기반의 정보 이론적 척도인 객체 및 속성 관계 엔트로피는 불완전 정보시스템을 완전 정보시스템으로 전환시킴으로 정보의 정확성을 향상시키는 유용한 방법이 된다. 그리고, 객체 및 속성 관계 엔트로피는 정보 시스템의 조건속성과 결정속성에 포함될 수 있는 null 및 불일치 값에 대한 대체 가능한 값을 결정하여 완전 정보시스템으로 구성할 수 있으며, 속성의 동치 관계 클래스에 대한 각각의 확률을 계산해야 하는 Beaubouef의 러프 엔트로피보다 계산 방법을 단순화시켰다. 또한, 객체 및 속성 관계 엔트로피는 지식 베이스를 구성하기 위한 추론 규칙의 정확성을 향상시키기 위한 전 처리 과정으로 수행될 수 있다.

참 고 문 헌

[1] Beaubouef, T., Petry, F. E. and Arora, G., "Information-theoretic measures of uncertainty for rough sets and rough relational databases," *Information Science*, Vol. 109, No. 1-4, pp. 185-195, 1998.

[2] Kanal, L. and Lemmer, J., *Uncertainty in Artificial Intelligence*, Amsterdam: North Holland, 1986.

[3] Kryszkiewicz, M., "Rules in incomplete information systems," *Information Science*, Vol. 113, No. 3-4, pp. 271-292, 1999.

[4] Lin, T. Y. and Cercone, N.(eds), *Rough Sets and Data Mining-Analysis of Imperfect Data*, Boston: Kluwer Academic publishers, 1997.

[5] Pawlak, Z., "Rough Sets," *International Journal of Information and Computer Sciences*, Vol. 11, No. 5, pp. 341-356, 1982.

[6] Pawlak, Z., *Rough Sets - Theoretical Aspects of Reasoning about Data*, Kluwer, 1991.

[7] Pawlak, Z., "Rough Set Theory and Its Applications to Data Analysis," *Cybernetics and Systems: An International Journal*, pp. 661-688, 1998.

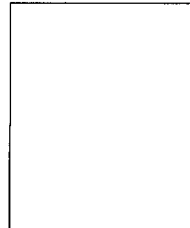
[8] Shannon, C. L., "The mathematical theory of communication," *Bell System Technical Journal*, Vol. 27, 1948.

[9] Słowiński, R. and Stefanowski, J., "Rough classification in incomplete information systems," *Mathematical and Compute. Modelling*, Vol. 12, No. 10-11, pp. 1347- 1357, 1989.

[10] 정구범, 김두완, 정환묵, "러프집합에 의한 불완전 데이터의 처리에 관한 연구," 한국 퍼지 및 지능시스템학회 '98 춘계학술대회 학술발표 논문집, pp. 11-15, 1998.

[11] 정구범, 정환묵, "러프집합을 이용한 정보시스템에서의 불완전 데이터 처리에 관한 연구," 퍼지 및 지능시스템학회 논문지, Vol. 9, No. 3, 1999.

김 국 보

대구 가톨릭대학교 박사
현재 대진대학교 컴퓨터공학과 교수
관심분야 : 시스템 엔지니어링, 인공지능

정 구 범

대구 가톨릭대학교 박사
현재 상주대학교 컴퓨터공학부 교수
관심분야 : 인공지능, 전자상거래

박 경 옥

연세대학교 박사
현재 삼성전자(주) 디지털프린팅 사업부 선임연구원
관심분야 : 통계학, 퍼지집합